Classifying skin cancer within a dermatoscopic image dataset

Dawn Finzi and Mona Rosenke

I. Introduction

Skin cancer is the most common form of cancer in the United States with one estimate putting the number of cases at 5.4 million per year [1]. While the large majority of these cases are non-lethal, early and accurate diagnosis is crucial for improving outcomes, especially for more aggressive skin cancers such as melanomas. In this project, we will be attempting to classify seven categories of pigmented lesions from dermatoscopic images, a task that closely matches the real-world clinical scenario, as all of the main categories of pigmented lesions are represented. We propose a two-pronged approach to this problem, which will be covered in II. *C. Planned Implementation*.

II. Methodology

A. Dataset

The dataset we will be using is an image dataset from Kaggle consisting of 10015 dermatoscopic images of pigmented lesions (Skin Cancer MNIST: HAM10000). Cases include seven main categories of pigmented lesions

- Actinic keratoses and intraepithelial carcinoma / Bowen's disease
- Basal cell carcinoma
- Benign keratosis-like lesions
- Dermatofibroma
- Melanoma
- Melanocytic nevi
- Vascular lesions

where three of the seven are precancerous or cancerous (actinic keratoses, basal cell carcinoma and melanoma) and the other four are largely benign (keratosis-like lesions, dermatofibroma, melanocytic nevi and vascular lesions).

All of the images are labeled based on histopathology (50% of cases), follow-up examination, expert consensus or confirmation by in-vivo confocal microscopy.

B. Concrete Example of Input and Output

For inputs, the system will take dermatoscopic images such as these (Fig. 1) and return one of seven labels corresponding to the seven main categories of pigmented lesions outlined above.

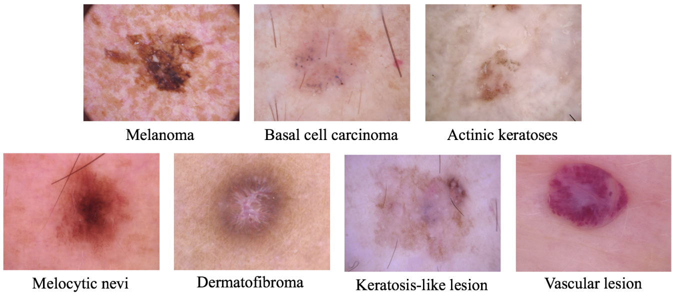

Fig. 1. Example input images for each category and their corresponding labels. Cancerous lesions are on the top row and benign lesions are on the bottom row

C. Planned Implementation

We propose a two-pronged approach to this task:

1) First, we will attempt to classify this dataset using a convolutional neural network. Specifically, we will train CNNs based on the ResNet50 and VGG16 networks, and assess performance accuracy on held-out test data for these competing architectures.
2) Secondly, we are interested in the image features humans use to make these classifications and if we could incorporate some of these metrics to simplify the computational power required for classification. We hope that the CNNs in Step 1 will perform well on the classification problem but we recognize that training and testing such large networks is unrealistic in the average clinical environment. We hope to be able to simplify the problem by reasoning about what heuristics dermatologists use to discriminate between the different lesion types and then encorporate these features more explicitly into our models.

D. Evaluation Metric

We will evaluate performance using two metrics

1) Classification accuracy: Percent correct on a held out test set, measured as model label prediction for each image vs. the ground truth labels from histopathology, expert consensus etc.
2) Resource utilization: We will test all models on the same GPUs (hopefully using Google Cloud computing) and then measure the inference time on a test image for each model. We presume that all models will only need to be trained once, so this is the most clinically applicable measure. We intend for this measure to approximate the amount of time a doctor will wait for results on a given dematoscopic image (scaled for decreased processing power).

III. Baseline

For our baselines, we used the RBG values of 28 pixel x 28 pixel versions of the images and implemented three out of the box models from sci-kit learn. For all three classifiers we used 80% of the data as training data (8012 images) and 20% as test data (2003 images). We first implemented a linear support vector classifier (LinearSVC) which classified test data with 60.0% accuracy. We then tried logistic regression using the ”lbfgs” solver for our multiclass problem. This performed better with 69.7% accuracy on the test data. Finally, we used the default scikit-learn instantiation of k-nearest neighbors classifier with five neighbors. The algorithm also performed decently on the test data, with classification accuracy at 70.1%. Note that chance on this dataset is 14%, which means that fairly decent performance can be achieved even with these more basic models. Thus, these baselines will provide a useful benchmark against which to measure our performance.

IV. Oracle

For our oracle, we initially thought to recruit a board-certified dermatologist who frequently diagnoses and treats melanomas and other skin cancer. We provided the dermatologist with 100 of the images and asked him to label the dematoscopic images based on the seven possible label categories. However, our proposed oracle only achieved 35% accuracy! While he did manage 89% accuracy when diagnosing the lesions as cancerous or benign, his 7-label categorization was far below our baseline performance. We speculate that this may be in part due to many of these lesions needing to be confirmed by histopathology, and also as the dermatologist we were able to recruit does not use dermatoscopy in his daily practice. Regardless, we were forced to explore other options for our oracle.

We then looked at the literature for recent state-of-the-art work that we could use as a benchmark. We discovered that an ISIC Challenge was held for this dataset in the summer of 2018 [3] and the winner of the challenge managed to achieve 89% on test data [4]. We will use this performance as our oracle. This approach uses an ensemble of over 19 models, including many convolution neural networks (CNN), rendering it essentially impossible to use in a real-world clinical setting. Thus this will serve well as an upper bound on our attempt to instantiate a more practical classifier.

V. Challenges

We anticipate that our biggest challenge will be successfully identifying what features dermatologists focus on in order to make diagnoses and then translating that into a tangible input. We hope to be able to survey a small sample of dermatologists but recognize that this will pose a challenge regardless.

VI. Related Work

Using artificial intelligence to diagnose skin cancer has been a hot topic lately with a highly publicized study [5] pitting ’man against machine’ published in the August 2018 issue of Annals of Oncology [6]. In practice, this study compared the performance of a CNN to 58 dermatologists on classifying dermatoscopic melanoma. Using Google’s Inception V4 CNN architecture, the authors showed that the neural network model outperformed most dermatologists as measured by sensitivity, specificity and area under the curve (AUC) of receiver operating characteristics (ROC) for classification. Here the classification problem was binary (melanoma vs. non-melanoma), allowing for computation of more sensitive performance metrics based on signal detection theory. However, this research further supports the idea that CNNs can acheive high performance on skin cancer classification problems, as well as illustrating the broad interest in methods for solving such a diagnostic problem. We also plan to research the Inception V4 CNN architecture and consider whether aspects should be incorporated into our models.

A second main source of related work stems from the ISIC Challenge [3]. Task 3 of this challenge specifically looked at 7-class classification accuracy on this very dataset. As well as using the winner of the contest as our oracle performance, we also plan to look at the leaderboard manuscripts in order refine our choice of models when implementing approach #1.

References

[1] https://www.cancer.org

[2] Nithin D. Reddy, *Classification of Dermoscopy Images using Deep Learning.* arXiv preprint arXiv:1808.01607, 2018

[3] https://challenge2018.isic-archive.com/leaderboards/

[4] Aleksey Nozdryn-Plotnicki, Jordan Yap, and William Yolland, *Ensembling Convolutional Neural Networks for Skin Cancer Classification.*

[5] https://www.usnews.com/news/health-care-news/articles/2018-05-28/artificial-intelligence-beats-dermatologists-at-diagnosing-skin-cancer

[6] Haenssle HA, Fink C, Schneiderbauer R, et al., *Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists.* Annals of Oncology, 2018.